Foundational Scientific Methodology for Neural Networks

Modern neural network research has overwhelmingly adopted an empirical scientific methodology: propose an architectural modification, train models, compare statistical performance, and retain modifications that produce measurable improvements. This approach has been extraordinarily successful in producing increasingly capable systems. However, statistical validation is only one component of the scientific method. Equally important is the search for explanatory mechanisms that reveal *why* a system behaves as it does.

A mature scientific discipline does not stop once a phenomenon is shown to work. Instead, it seeks multiple complementary explanations of the same underlying mechanism until the essential behavioral principles become clear. The goal is not merely prediction, but understanding.

This distinction can be illustrated by the role of the weighted sum in artificial neurons. It is commonly presented simply as a linear combination of inputs. Yet the same mathematical operation can be understood from several equally valid perspectives:

- as a matched filter that measures similarity to a learned pattern,
- as an associative memory retrieving stored correlations,
- as a projection onto a learned feature direction,
- as evidence accumulation,
- as a geometric operation defining a hyperplane.

Each viewpoint emphasizes different aspects of the same computation. None is uniquely "correct"; together they produce a richer understanding than any single interpretation.

The same applies to nonlinear activation functions. The Rectified Linear Unit (ReLU), for example, is typically introduced as a computationally convenient nonlinearity that improves optimization. While true, this explanation captures only one aspect of its behavior. ReLU may also be viewed as a local gating mechanism, a binary switch controlling information flow, a context-dependent selector, a partition of the input space into distinct computational regions, or a mechanism for inducing sparse representations. Each interpretation highlights different functional properties that become important in different theoretical contexts.

The scientific value lies not in choosing one interpretation, but in examining the mechanism from many perspectives until stable behavioral principles emerge.

Historically, neural network research has often favored a single dominant conceptual narrative inherited from its origins. The artificial neuron was primarily interpreted as an abstraction of a biological neuron, and many subsequent explanations were built upon that historical foundation.

As empirical successes accumulated, this narrative became increasingly detached from the mechanisms that appeared to govern the behavior of modern deep networks. Components were retained because they improved benchmark performance rather than because they fit within a coherent explanatory framework.

This is not a criticism of empirical methodology itself. Statistical comparison is an indispensable part of science. However, statistical validation answers only the question of *whether* a modification improves performance under particular conditions. It does not by itself explain *what computational role the modification performs*, *which aspects are essential*, or *whether the same underlying mechanism appears elsewhere in the system under different forms*.

Foundational scientific methodology requires repeated re-examination of basic mechanisms from multiple conceptual viewpoints. As understanding develops, explanations should become simpler, more unified, and more predictive. Rather than accumulating historical assumptions, the field should continuously refine its conceptual models to reflect the mechanisms actually responsible for observed behavior.

Viewed in this way, neural networks are not merely collections of empirically successful components. They are computational systems whose elementary operations can often be understood through multiple complementary abstractions. Progress toward a genuine theory of neural computation will likely depend less on discovering new architectural components than on developing richer mechanistic descriptions of the operations already present. Statistical evidence identifies phenomena worthy of explanation; foundational science seeks the explanatory principles that unify them.